



Nicht-kanonische Sprache parsen

Falko: Fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache



Arts & Humanities
Research Council



Falko



Wie können nicht-kanonische Sprachdaten für syntaktische Fragestellungen aufbereitet werden?

Nicht-kanonische Sprachdaten (Hirschmann et al. erscheint)

- werden von Standardgrammatiken nicht beschrieben

Lernertext:

Es **ist diese Gesetzte** die wichtig für die Gesellschaft **ist** sie von Gefahren zu schützen. (falko:sa009_2006_09_L2v2.2)

- vergleichbares Problem in gesprochener Sprache

gesprochene Sprache:

"beim Motorrad ... **mm gehs/** fährst du **unter das Motorrad** ...**durch** .. nach links " (hamatac:MT_091209_Ali)

Zwei mögliche Lösungen

- spezifische Grammatik für jede Varietät schreiben

→ **Nachteil:** - Varietäten können nur noch qualitativ unterschieden werden.

- Erstellung einer kanonischen Entsprechung

(Hirschmann et al. 2007)

→ **Vorteil:** - Abweichungen lassen sich auch quantitativ auswerten.
- Standard-Tools (Tagger, Parser etc.) können genutzt werden

Zielhypothesen (in Falko)

- Mit minimalen Korrekturen wird ein vom Parser verarbeitbarer Input erzeugt (ZH1). (Lüdeling et al 2008)
- Abweichungen vom der ZH1 vom Lernertext werden automatisch annotiert. (Reznicek et al. 2010)

LT	Es	ist	diese	Gesetzte		die	wichtig
ZH1	Es	ist	dieses	Gesetz	,	das	wichtig
ZH1Diff			CHA	CHA	INS	CHA	

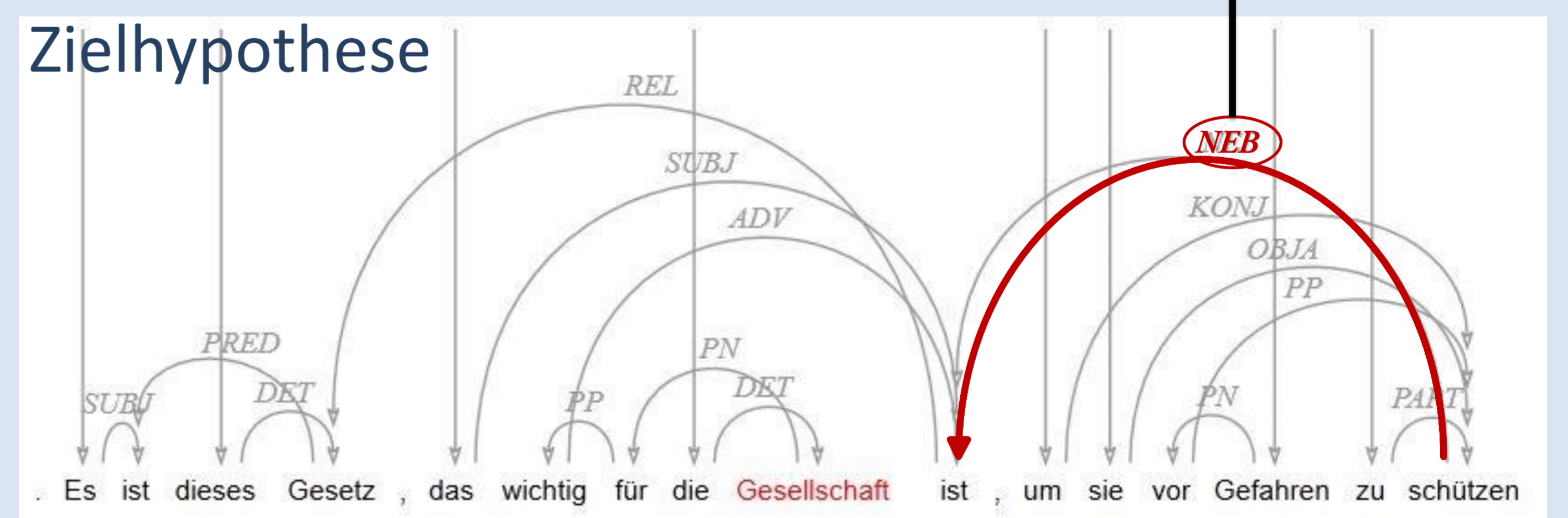
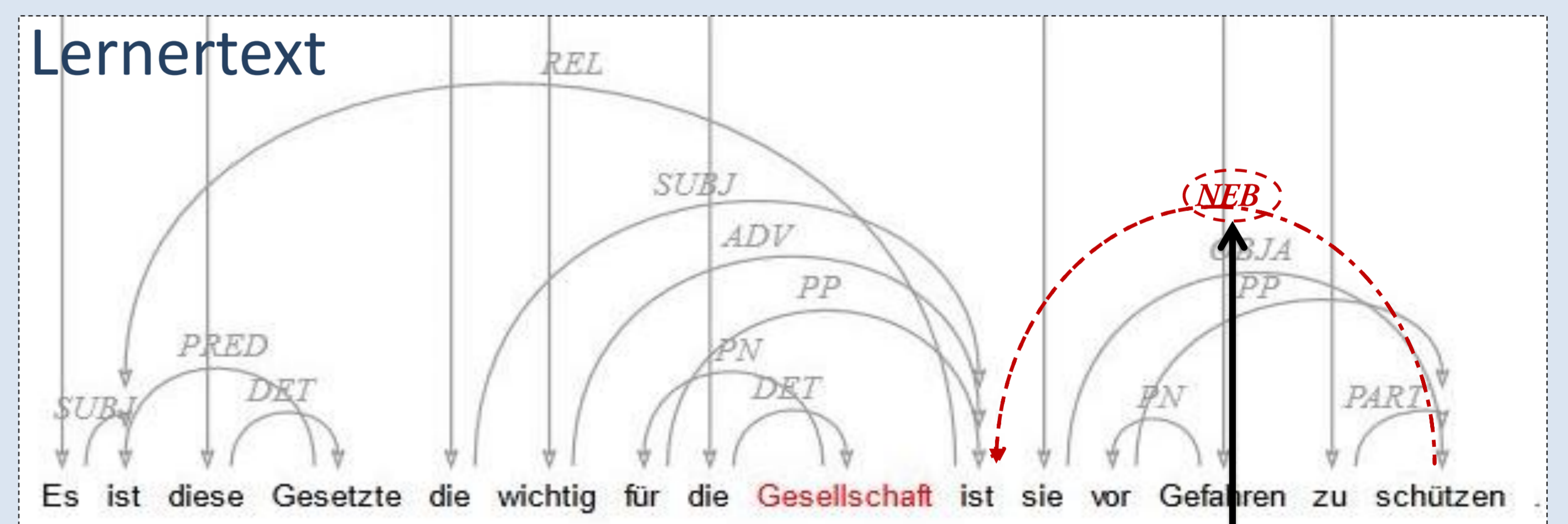
Tagging

- ZH1 erlaubt als Volltext automatische Annotation
 - pos, lemma, morph (rfTagger: Schmid/Laws 2008)

ZH1	Es	ist	dieses	Gesetz	,	das	wichtig
ZH1pos	PPER	VAFIN	PDAT	NN	\$,	PRELS	ADJD
ZH1lemma	es	sein	dies	Gesetz	,	d	wichtig
ZH1morph	3.Nom. Sg.Neut	3.Sg. Pres. Ind	Nom.Sg. Neut	Nom.Sg. Neut		Acc.Sg. Neut	

Parsing

- Annotation der syntaktischen Relationen und Funktionen erlaubt oberflächenferne Analysen. (Hirschmann et al. 2011)
- Abhängigkeiten (MALT: Nivre et al. 2007)
 - grammatische Funktionen
- PROBLEM:** Für den Beispielsatz findet der Parser keine geeignete Lösung.
 - Finalsatz wird nicht verbunden
- LÖSUNG:** Der Pars der Zielhypothese macht die finale Relation auffindbar.
 - Relation kann auf Lernertext übertragen werden.



Suche (Beispiel)

- Grammatische Funktionen können gezielt in ANNIS2 (Zeldes et al. 2009) gesucht werden:

node & node & #1->dep[func="NEB"]#2

beliebiger
Knoten

beliebiger
Knoten

Nebensatz

Falko@ANNIS2: <http://korpling.german.hu-berlin.de/falko-suche>

Ausblick

- Kombination einzelner Beschreibungsebenen
 - Abhängigkeiten (MALT: Nivre et al. 2007)
 - Topologische Felder (Stanford: Rafferty et al. 2008)
 - Minimale Chunks (Stanford)
- unkomplizierte manuelle Korrektur (vakyartha)
- transparente Suche (ANNIS2)

Literatur

Gerdes, Kim; vakyartha, <http://arborator.ilgga.fr/vakyartha/>, 09.03.2012.
Hirschmann, Hagen; Doolittle, Seanna; Lüdeling, Anke (2007): Syntactic Annotation of Non-Canonical Linguistic Structures. In: Proceedings of Corpus Linguistics 2007. Birmingham.
Hirschmann, Hagen; Lüdeling, Anke; Rehbein, Ines; Reznicek, Marc; Zeldes, Amir (erscheint): Underuse of Syntactic Categories in Falko. A Case Study on Modification. In: 20 years of learner corpus research. looking back, moving ahead (LCR2011). Learner Corpus Research 2011. Louvain-la-Neuve, Belgium, 15-17 September 2011.
Lüdeling, Anke; Doolittle, Seanna; Hirschmann, Hagen; Schmidt, Karin; Walter, Maik (2008): Das Lernerkorpus Falko. In: DAF 45 (2), S. 67-73.

Nivre, Joakim; Nilsson, Jens; Hall, Johan; Chanev, Atanas; Eryigit, Gülsen; Kübler, Sandra et al. (2007): MaltParser. A Language-Independent System for Data-Driven Dependency Parsing. In: *Natural Language Engineering* 13 (1), S. 1-41. <http://w3.msi.vxu.se/~nivre/papers/nle07.pdf>, 09.03.2012.
Rafferty, Anna; Manning, Christopher D. (2008): Parsing Three German Treebanks. Lexicalized and Unlexicalized Baselines. In: ACL Workshop on Parsing German.
Reznicek, Marc; Walter, Maik; Schmidt, Karin; Lüdeling, Anke; Hirschmann, Hagen; Krummes, Cedric; Andreas, Torsten (2010): Das Falko-Handbuch. Korpusaufbau und Annotationen. Version 1.01. <http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko>, 09.03.2012.
Schmid, Helmut; Laws, Florian (2008): Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In: Donia Scott (Hg.): *Coling 2008*, S. 777-784.
Zeldes, Amir; Ritz, Julia; Lüdeling, Anke; Chiaros, Christian (2009): ANNIS. A Search Tool for Multi-Layer Annotated Corpora. In: Proceedings of Corpus Linguistics 2009, Liverpool, July 20-23, 2009.